RADHEY GUPTA, McMaster University

A Signature Protein Based In silico Microbial Identification Tool Using Next Generation Sequence Data

Rapid and reliable interrogation of the next generation sequence (NGS) data for the presence or absence of different organisms poses a significant challenge which limits the routine use of NGS techniques for clinical diagnostic and metagenomic investigations. A new approach/tool is described here based on Conserved Signature Proteins (CSPs), which are sets of proteins that are uniquely found in specific groups of organisms, for rapid interrogation of the NGS data for the presence or absence of different organisms. A large database of validated CSPs has been created that are specific for different prokaryotic and some eukaryotic organisms at multiple taxonomic levels ranging from phylum to species/strain levels. All significant blast hits for these CSPs are for the indicated group(s) of organisms. Due to the predicted presence of these CSPs in the indicated groups of organisms, Blast searches with their sequences (amino acid or nucleotide) provide a highly specific mean for rapidly and reliably determining the presence or absence of organisms from these groups in the metagenomic sequences. Using these CSPs, an in silico Web-based Microbial Identification Tool has been developed for rapidly determining the presence or absence of either specific organisms, or comprehensive taxonomic profiling of different organisms, in metagenomic sequences.